

# **SYSTEM AND METHOD FOR IDENTIFYING LANGUAGE ATTRIBUTES THROUGH PROBABILISTIC ANALYSIS**

## **Abstract**

A system and method for identifying language attributes through  
5 probabilistic analysis is described. A set of language classes and a plurality of  
training documents are defined. Each language class identifies a language and a  
character set encoding. Occurrences of one or more document properties within  
each training document are evaluated. For each language class, a probability for  
the document properties set conditioned on the occurrence of the language class is  
10 calculated. Byte occurrences within each training document are evaluated. For  
each language class, a probability for the byte occurrences conditioned on the  
occurrence of the language class is calculated.